Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology on the Web

LF. Soualmia[1,2] , C. Barry[3] and SJ. Darmoni[1,2]

[1]CISMeF & L@STICS, Rouen University Hospital, 76031 Rouen, France
{lina.soualmia, stefan.darmoni}@chu-rouen.fr
[2]PSI FRE CNRS 2645, INSA-Rouen, 76131 Mont-Saint Aignan, France
[3]LaRIA, Picardie University, 80000 Amiens, France
barry@laria.u-picardie.fr

Abstract. This paper deals with the problem of information retrieval on the Web and present the CISMeF project (acronym of Catalogue and Index of French-speaking Medical Sites). Information retrieval in the CISMeF catalogue is done with a terminology that is similar to ontology of medical domain and a set of metadata. This allows us to place the project at an overlap between the present Web, which is informal, and the forthcoming Semantic Web. We also describe an ongoing work, which consists of applying three knowledge-based methods in order to enhance information retrieval.

1 Introduction

Nowadays the problematic is *intelligent information retrieval* on the Web. The Semantic Web [1] is an infrastructure that has to be built. It aims at creating a web where information semantics are represented in a form that can be understood by human as well as machines in order to enable computers and people to work in co-operation. One of its advantages is to bring sufficient information on the resources, by adding annotations in the form of *metadata* and to describe formally and significantly their content according to an *ontology*. Ontologies are considered to be powerful tools to lift ambiguity by providing a controlled vocabulary of terms and some specification of their meaning and are very useful for interoperability and for browsing and searching. Metadata describe Web information resources enhancing information retrieval.

In this paper we present the CISMeF[1] project [2] (acronym of Catalogue and Index of French-speaking Medical Sites) developed since 1995. The objective of CISMeF is to help health professionals, as well as students and the general public, during their search for electronic health information. The CISMeF catalogue describes and indexes a large number of health information resources (*n*=11,504). A resource can be a Web site, Web pages, documents, reports and teaching material: any support that may contain health information. The resources are selected according to strict criteria by the librarian team and are indexed according to a methodology. The resources

[1] http://www.chu-rouen.fr/cismef/

indexed in the CISMeF catalogue are described according to a terminology that is similar to an ontology of the medical domain, and a set of metadata elements. This structure enables us to place the project at an overlap between the present informal Web, mainly composed by HTML pages, and the forthcoming Semantic Web. We also describe in this paper an ongoing work which consists of applying three knowledge-based methods (natural language processing, knowledge discovery in databases and reasoning on ontologies) to enhance information retrieval into CISMeF.

2 Towards a Medical Semantic Web

Metadata is data about data and specifically in the context of the Web, it is data that describe Web resources. When properly implemented, metadata can enhance information retrieval. In CISMeF several sets of metadata elements are used. The resource indexed are described the Dublin Core (DC) elements set [3] (e.g. *author, date*). DC is not a complete solution, it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF [2](e.g. *institution, target public*). Two additional fields are in the resources intended for the health professionals: indication of the *evidence-based medicine* and the *method* used to determine it. In the teaching resources eleven elements of the IEEE 1484 LOM (Learning Object Metadata) “Educational” category are added. The metadata format was the HTML language in 1995. Since December 2002, the format used is RDF, a Semantic Web language, within the ongoing MedCIRCLE project [4], developed to qualify health information quality.

The catalogue resources are indexed according to the CISMeF terminology, which is based on the MeSH [5] concepts and its French translation. We have not used the UMLS [6] because there is no available French translation. Approximately 22,000 keywords (e.g. hepatitis) and 84 qualifiers (e.g. complications) compose the MeSH thesaurus, in its 2003 version. These concepts are organized into hierarchies going from the most general on the top to the most specific in the bottom of the hierarchy. The qualifiers, organized into hierarchies, specify which particular aspect of a keyword is addressed. The keywords and the qualifiers that are dispersed in several trees but are semantically related in CISMeF are gathered according to *metaterms* (*n*=66). They concern medical specialties. In addition, a hierarchy of *resource types* (*n*=127) describes the nature of the resource (e.g. *clinical guidelines*). The metaterms and resource types enhance information retrieval into the catalogue when searching for “*guidelines in cardiology*”, where *cardiology* is a metaterm and *guidelines* is a resource type, which is not possible using the MeSH thesaurus.

The CISMeF terminology has the same structure as a terminological ontology [7]. The vocabulary describes major terms of the medical domain and is well known by the librarians and the health professional. Each concept has a *preferred term* to express it in natural language, a set of properties, a natural language definition that allows to differentiate it from the concepts it subsumes and those that subsume it, a set of synonyms and a set of rules and constraints.

3 Enhancing Information Retrieval

The submitted queries over the search engine are seldom matched to the terms of the ontology. We have extracted and analyzed 1,552,776 queries of the http server log and their associated number of answers (between the 08/15/02 and the 02/06/03). 892,591 queries (58.62%) were submitted via the free text search interface [2] and 365,688 (40.97%) had no answer.

3.1 Natural Language Processing

We apply here a morphological analysis of the queries. A preliminary work [8] showed that using morphological knowledge enhance information retrieval. The proposed algorithm consists in correcting the user query by eliminating stop words (*the, and, when*) and replacing each word of the query by a disjunction of all the terms of its morphological family. A morphological family of a term is composed by its *inflexions* {*accident, accidents*} and *derivations* {*probability, probabilistic*}. If the user query is "*interaction between the drugs*" it will be replaced by the MeSH term "*drug interactions*". There is not yet an available French Medical Lexicon, such as the Specialist Lexicon of the UMLS, so we have used a terminological resource Lexique [9] that is not related to the medical domain. Nevertheless, it allowed us to obtain 31,016 derived terms that match exactly 1,292 CISMeF terms.

Table 1. Structure of the terms used for indexing the resources.

Number of words	Keywords	Qualifiers	Resource Types	Terms
1	1 437	55	28	1 520
2 to 7	2 516	24	99	2 639
TOTAL	**3953**	**79**	**127**	**4 159**

Table 2. Matching the vocabulary

	Keywords	Qualifiers	Resource Types	Terms
Nb terms matched	1 207	55	28	1 292
1 word matching	83.99%	100%	100%	85%
Semi-matching	78.57%	79.74%	39.37%	77.59%
Total matching	30.53%	69.62%	22.04%	31.06%

The analysis of the other terms composed by 2 or more words showed that 1,935 terms (1,899 keywords; 8 qualifiers; 22 resource types) are *semi-matched*. A term is *semi-matched* if at least one of the words that compose it is matched. In addition to morphological knowledge, semantic knowledge is necessary, for example *heart* and *cardiac* are semantically related. A set of synonyms has been created with the collaboration of several patients associations and we are currently analyzing the user queries to complete it.

3.2 Knowledge Discovery in Database

We want to discover "new" knowledge from the CISMeF database to exploit it in the process of information retrieval. We apply a Data Mining technique called *Association Rules* to extract interesting associations, previously unknown, from the database. A Boolean association rule AR is expressed as:

$$\mathrm{AR}: i_1 \wedge i_2 \wedge \ldots \wedge i_j \Rightarrow i_{j+1} \wedge \ldots \wedge i_n \qquad (1)$$

This formula states that if an object has the items $\{i_1, i_2 \ldots, i_j\}$ it tends also to have the items $\{i_{j+1}, \ldots, i_n\}$. The AR *support* represents its utility. This measure corresponds to the proportion of objects which contains at the same time the rule antecedent and consequent. The AR *confidence* represents its precision. This measure corresponds to the proportion of objects that contains the consequent rule among those containing the antecedent. The extraction context is the triplet C= (O, I, R) where O is the set of objects, I the set of all the items and R a binary relation between O and I. The objects are the annotations used to describe the indexed resources. The relation R represents the indexing relation between an object and an item. We first consider two cases for the items: I_1={Keywords} and I_2={(Keywords/Qualifiers)}. An itemset is frequent in the context C if its support is higher than the minimal threshold initially fixed. We use the A-Close algorithm [10], which deduces bases for association rules. We have tested our algorithm on few examples. The first step of the algorithm allowed us to find for example the following rules: *Hepatitis C* ⇒ *AIDS* with support=14 for I_1 and *AIDS/prevention and control* ⇒ *condom* with support=6 for I_2. The second step is to extract all the other association rules and to apply them in the information retrieval process by a query expansion.

3.3 Reasoning on Ontologies

In order to complete the CISMeF ontology with more refined links between concepts, we have decided to exploit the UMLS Semantic Network, which is composed by medical concepts and semantic relations between concepts. They take the form of *Complications (Hepatitis, Liver Cirrhosis)* denoting that the concept *Hepatitis* is related to the concept *Liver Cirrhosis* by the relation *Complications*. These relations correspond to the MeSH qualifiers and the concepts correspond to the MeSH keywords. These relations won't be used to annotate the resources but they will be converted into inference rules enriching by that the ontology by other links between concepts. In our example, the only one information available from the ontology is that the concepts *Hepatitis* and *Liver Cirrhosis* are subsumed by the concept *Liver Diseases*. In order to enable content reasoning over the resources, we will convert a part of the CISMeF ontology into RDF Schema by transforming keywords and resource types into *concepts* and the qualifiers into *roles* (or relations). The resources will be transformed into RDF according to the CISMeF RDF Schema. RDFS doesn't include reasoning mechanisms such as those included in the Description Logics Systems but unlike RDFS, the query languages for the other ontology standards are still ongoing. Writing inference rules with RDFS is possible with TRIPLE [11]: it has been developed for knowledge-based intelligent information retrieval. It enables to carry out complex reasoning on RDF resources that represent the concepts' instances. In our case, for example, if a resource R is an instance of *Hepatitis/Complications* and the user is searching for resources related to *Liver Cirrhosis*, the system would infer that the resource R is also an answer to the query. We will use the TRIPLE query engine to carry out higher level queries over the CISMeF catalogue.

4 Conclusion and Future Work

We have discussed in this paper the problems of information retrieval on the Web. We have presented particular aspects of the CISMeF project. We have also proposed different methods to enhance information retrieval. The natural language processing is used to build morphological knowledge base. Data Mining enables association rules discovery between concepts. Finally, reasoning on ontologies will offer a higher level for the ontology (consistency and coherence checking, exploitation of the Semantic Network of the UMLS) and for information retrieval. To our knowledge, no existing work has combined these methods in order to enhance information retrieval. The next step of this study is to evaluate the contribution of each method separately and conjointly: we will apply an automatic and an interactive query expansion over the users' queries. The evaluation on a real scale will allow us to deduce a process, according to the type of the query, to apply a method with a particular order.

References

1. Berners-Lee, T., Heudler, J. and Lassila, O. (2001). The Semantic Web. Scientific American, p.35-43.
2. Darmoni, SJ., Thirion, B., Leroy, JP. et al. (2001). A Search Tool based on 'Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. Medical Informatics & the Internet in Medicine, 26 (3) :165-178.
3. Baker, T.(2000) A Grammar of Dublin Core. Digital-Library Magazine, vol 6 n°10.
4. Mayer, MA., Darmoni, SJ., Fiene, M. et al. (2003) MedCIRCLE - Modeling a Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information on the Semantic World-Wide-Web. Medical Informatics Europe, p.667-672.
5. Nelson, SJ., Johnson, WD., and Humphreys, BL. (2001) Relationships in Medical Subject Headings. Bean and Green (eds). Kluwer Academic Publishers, 171-184.
6. Lindberg, DAB, Humphreys, BL and McCray, AT. (1993) The Unified Medical Language System. Methods of Information in Medicine, 32 (4):281-291.
7. Sowa, JF. (2000) Ontology, Metadata and Semiotics. Lecture Notes in AI #1867, Springer Verlag, p.55-81.
8. Zweigenbaum, P., Darmoni, SJ. and Grabar, N. (2001) The Contribution of Morphological Knowledge to French MeSH Mapping for Information Retrieval. JAMIA 8:796-800.
9. New, B., Pallier, C., Ferrand, L. and Matos R. (2001) Une Base de Données Lexicales du Français Contemporain sur Internet: LEXIQUE, L'Année Psychologique, 447-462.
10. Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. (1999) Efficient Mining of Association Rules Using Closed Itemset Lattices. Information Systems, 24(1):25-46.
11. Sintek, M. and Decker, S. (2001) TRIPLE- An RDF Query, Inference and Transformation Language. Proceedings of Deductive Databases and Knowledge Management Workshop.